The effect of sequence mismatches on binding affinity and endonuclease activity are decoupled throughout the Cas9 binding site

Liyang Zhang[1], H. Tomas Rube[2,3], Harmen J. Bussemaker[2,3], and Miles A. Pufall[1]

[1]Department of Biochemistry, Carver College of Medicine, University of Iowa; [2]Department of Biological Sciences, Columbia University, New York, NY 10027; [3]Department of Systems Biology, Columbia University, New York, NY 10032

Corresponding author: miles-pufall@uiowa.edu (M.A.P.)

ABSTRACT

The CRISPR-Cas9 system is a powerful genomic tool. Although targeted to complementary genomic sequences by a guide RNA (gRNA), Cas9 tolerates gRNA:DNA mismatches and cleaves off-target sites. How mismatches quantitatively affect Cas9 binding and cutting is not understood. Using *SelexGLM* to construct a comprehensive model for DNA-binding specificity, we observed that 13-bp of complementarity in the PAM-proximal DNA contributes to affinity. We then adapted *Spec-seq* and developed *SEAM-seq* to systematically compare the impact of gRNA:DNA mismatches on affinity and endonuclease activity, respectively. Though most often coupled, these simple and accessible experiments identified sometimes opposing effects for mismatches on DNA-binding and cutting. In the PAM-distal region mismatches decreased activity but not affinity, whereas in the PAM-proximal region some reduced-affinity mismatches enhanced activity. This mismatch-activation was particularly evident where the gRNA:DNA duplex bends. We developed integrative models from these measurements that estimate catalytic efficiency and can be used to predict off-target cleavage.

INTRODUCTION

The CRISPR-Cas9 system is driving the next wave of genetic research. The flexibility and relative ease of using a programmable guide RNA (gRNA) to target Cas9 to specific complementary DNA sequences has made genomic editing readily accessible (1, 2). Further, catalytically deactivated Cas9 (dCas9) has been adapted to manipulate gene expression and chromatin state, visualize chromosomal loci, and other applications (3-6). Despite active research in the area, developing rules that predict the effect of mismatches on off-target binding and endonuclease activity have had limited success, impeding broader use of Cas9.

The Cas9 ribonucleoprotein (Cas9-RNP), composed of the protein and annealed CRISPR RNA (crRNA) and trans-acting crRNAs (tracrRNA), cuts DNA in a two-step process. First the Cas9-RNP binds specific DNA sequences called protospacers (7). The Cas9 protein itself binds the sequence NGG called protospacer adjacent motif (PAM), and partially melts the DNA duplex. The gRNA, within the crRNA, then basepairs with the 20-nt protospacer DNA. The resulting RNA:DNA duplex form an R-loop that then basepairs in a sequential fashion from the PAM toward the distal region of the protospacer (8, 9). The resulting "zipping model" for target identification and binding has shown that mismatches between the gRNA and protospacer affect the kinetics and affinity of binding by interrupting zipping, with adjacent mismatches having a more than additive effect. A structural change in the Cas9 protein accompanies basepairing in the PAM-distal region, activating the endonuclease, and cutting both strands of DNA upstream of position 3 in the protospacer (7).

One consequence of the zipping model is that the gRNA:protospacer duplex can be divided into two functional domains. Basepairing within the seed, or PAM-proximal region (positions 1-10), contributes most strongly to targeting the Cas9-RNP to specific sequences. Base-pairing in the PAM-distal region (positions 11-20) does not contribute as strongly to affinity, but is important for enzymatic activity by inducing the conformational change that activates Cas9 protein (10). Thus, the current model is that the seed sequence targets Cas9-RNP to binding sequences whereas the PAM-distal region is important for the Cas9-RNP endonuclease activity (10, 11). However, the boundaries of each domain are not clear, nor are the quantitative effects of mismatches on binding, endonuclease activity, and overall enzyme efficiency within each of these regions.

The DNA-binding specificity of Cas9-RNPs has primarily been inferred from genomic localization in cells using chromatin immunoprecipitation followed by deep sequencing (*ChIP-seq*). In addition to the intended target, the Cas9-RNP binds from a handful to thousands of other sites in the genome (12, 13). These off-targets sites still retain elements of the PAM and matching protospacer, with a consensus sequence ranging from 5 to 13 bp of complementarity. Further, although generally mismatches within the PAM-distal region (positions 11-20) are better tolerated than mismatches within the seed region, the Cas9-RNP can still bind sites with some seed region mismatches. More recently, high-content kinetic studies of Cas9-RNP binding *in vitro* have shown that mismatches in the PAM-proximal region dissociate more readily than distal mismatches, but that distal mismatches can

accentuate the effect of proximal mismatches (14). Although these studies are consistent, they do not present a comprehensive model for the DNA-binding specificity of the Cas9-RNP.

In addition to off-target binding, off-target cleavage by Cas9-RNPs can be prevalent (15). Measuring the prevalence of *in vivo* off-target cleavage has been challenging. Detection of cleavage in cells relies on detection of the indels, mutations caused by non-homologous end joining (12, 15-17), or by tagging double-strand breaks to mark off-target sites (18, 19). Although improving in sensitivity, these techniques still rely on cellular repair mechanisms, which vary from cell to cell. Nonetheless, it is clear that mismatches throughout the PAM and full 20-bp protospacer affect cleavage efficiency. A comparison of activities in the same cells shows a remarkable lack of overlap between the most highly bound and cleaved sites (12). Further, recent reports posit that sites with little complementarity can be cleaved by the Cas9-RNP *in vivo*, and that off-target editing is in fact widespread (18). These findings suggest a distinction between binding and cleavage that is not fully appreciated.

In vitro measurements of off-target cleavage have confirmed that efficient cleavage requires both an intact NGG within the PAM and complementarity throughout the protospacer. Using randomized libraries or isolated genomic DNA, thousands of off-target cleavage sites are detectable (19). Analogous to their effect on binding, mismatches appear to have a more pronounced impact on cleavage efficiency the closer they are to the PAM, and some are more deleterious than others (2, 16, 20-22). Although this reinforces the model that endonuclease activity is governed by affinity in the seed region, the lack of correlation between genomic localization and where the enzyme cleaves the genome suggest that this model is lacking.

Thus, despite significant progress in the identification of off-target cleavage and binding of the Cas9-RNP, the impact of mismatches on overall enzyme efficiency is not known. Modeling enzyme efficiency requires measurement of affinity and activity under the same conditions. To do this, we developed an *in vitro* approach to systematically measure both in parallel. We first used a technique we recently developed, *SELEX-seq* followed by generalized linear model fitting (*SelexGLM*) (Zhang, in revision, (23), to generate a biophysical model of Cas9-RNP binding specificity over a large (>30bp) binding site. To separately quantify the effect of mismatches on endonuclease activity and DNA-binding of the Cas9-RNP, we developed a new technique, Sequence-specific endonuclease activity measurement by sequencing (*SEAM-seq*), adapted Specificity measured by sequencing (*Spec-seq*) (24, 25), and developed new mechanistic models of Cas9-RNP specificity. Together, these approaches allowed us to identify regions where the effects are different or even opposing.

MATERIAL AND METHODS

Recombinant Cas9 expression and purification. The plasmid encoding WT-Cas9 (pMJ915) was a gift from Jennifer Doudna (Addgene plasmid # 69090). Point mutations (D10A and H840A) were created by site-directed mutagenesis to express catalytic deactivated Cas9 (dCas9) using primer pMJ915_D10A and pMJ915_H840A (See **Table S1** for sequence). Both Cas9 and dCas9 were expressed and purified as previously described (26). Briefly, the Cas9 protein was overexpressed in Rosetta 2 DE3 cells, and sequentially purified using Ni^{2+} affinity, cation exchange and size exclusion chromatography. N-terminal His- and MBP-tags were removed by TEV protease prior to the cation exchange chromatography. The purified protein was dialyzed into storage buffer (20 mM HEPES-KCl, pH7.5, 150 mM KCl, 10% Glycerol, and 1 mM TCEP), concentrated by Amicon ultrafiltration device (30-kDa), aliquoted, flash-frozen in liquid nitrogen, and stored at -80ºC. The protein concentration was determined by Nanodrop using an extinction coefficient of 120,450 $M^{-1}cm^{-1}$.

Cas9-RNP assembly. The NanogSg3 single guide RNA was synthesized (Synthego), and dissolved in the Cas9 storage buffer to a final concentration of 10 µM. The sgRNA was refolded by first incubating at 70ºC for 5 minutes, and then gradually cold down to 25ºC over 30 minutes. $MgCl_2$ was added to a final concentration of 1 mM, followed by incubating at 50ºC for 5 minutes, gradually cold down to 25ºC over 30 minutes (27). Refolded sgRNA was aliquoted and stored at -80ºC. To assemble the Cas9-RNP, 3 µM of purified Cas9 was incubated with 3.6 µM refolded sgRNA at 37ºC for 10 minutes (20 mM HEPES-KOH, pH 7.5, 150 mM KCl, 10% Glycerol, 1 mM $MgCl_2$, and 1 mM TCEP). Freshly assembled Cas9-RNP was used for each assay.

SelexGLM. A DNA library with 30-bp randomized region (**Table S1)** was synthesized as single-stranded DNA by IDT using hand-mixing option. The double-stranded DNA library (dsDNA) was prepared by a Klenow extension reaction using Cy5-labeled TSSR1 primer (**Table S1**) that is complementary to the 3'-end of the ssDNA library. Briefly, a reaction containing 2.5 µM ssDNA template, 5 µM Cy5-labeled TSSR1, and 150 µM dNTP in NEB buffer 2 (10 mM TrisHCl, 50 mM NaCl, 10 mM $MgCl_2$, 1 mM DTT) was incubated at 94ºC for 3 minutes, and then slowly drop to 37ºC over 45 minutes. Klenow enzyme was then added to the reaction, and incubated at 37ºC for 1 hour. The enzyme was inactivated at 72ºC for 20 minutes, followed by gradually cooling down to 10ºC over 45 minutes. The dsDNA library was purified and concentrated (Qiagen MinElute). The concentration was measured by A260 on a Nanodrop, and then diluted to 4 µM in Qiagen buffer EB (10 mM Tris-HCl, pH 8.5).

SELEX-seq of dCas9-RNP was performed based on our previous protocol with slight modifications. Briefly, a 120 µl binding reaction containing 0.1 µM dCas9-RNP and 1 µM SELEX library (1:10 ratio) was incubated at 37ºC for 1 hour (20 mM HEPES-KOH, pH7.5, 150 mM KCl, 10% Glycerol, 5 mM $MgCl_2$, and 1 mM TCEP), then resolved on a 4~20% (79:1) gradient gel (1X TGM: 25 mM Tris-Base, 192 mM Glycine, 5 mM $MgCl_2$, pH 8.3) at room temperature. The monomeric dCas9-RNP:DNA complex was recovered by excision and electroelution (200V, 4˚C in pre-chilled 1X TGM buffer). The bound DNA was purified (Qiagen MinElute), and diluted to 190 µl (EB buffer). Quantitative PCR was performed to determine the optimal number of rounds of amplification before saturation (thus avoiding

many artifacts) using 1/190 µl recovered DNA (Zhang, in revision). The next library was then generated by amplifying the remaining recovered DNA (~180 µl) in a 9mL PCR using the maximal number of rounds. The new library was then generated by purifying the regenerated library (Qiagen MinElute) then diluting to 4 µM for next round of *SELEX*. Five rounds of *SELEX* were performed in total. The initial and final libraries (R0 and R5) were deep-sequenced on an Illumina HiSeq at a read depth of ~25-30 million reads per library.

The *SELEX-seq* data was processed by the *SelexGLM* and *SELEX* R packages as described previously (Zhang, in revision). Briefly, a Markov model of order 6 was constructed from the R0 probes using the selex.mm() function from the *SELEX* package, and an affinity table for k=18 was constructed using selex.affinities(). An initial position specific affinity matrix (PSAM) was constructed from the relative affinity of all single-base mutations of the optimal 18-mer ("NNNNAAGAWKGGAAGNGG"). The PSAM was then expanded to the desired size by adding 9 neutral columns on each side to estimate the specificity outside of PAM and protospacer, and used as a seed for *SelexGLM*. The subsequent iteration of *SelexGLM* contains two main steps. First, the current PSAM is used to compute the affinity of each 36-bp window on each DNA probe. If the highest affinity window is greater than 95% of the total affinity (sum of all windows), the probe is saved for generating new PSAM. Next, a logarithmic link function is used to fit the base identity of the optimal window on each probe against the probe counts normalized by the frequency in Round 0 library computed by a 6th-order Markov model. The regression coefficients are interpreted as the free energy differences ΔΔG, and used in the next round of iteration. This iterative process continues until the PSAM converges. The final model was plotted as motif logo using REDUCE Suite (28).

Spec-seq. Each individual *Spec-seq* library was ordered as ssDNA from IDT, and pooled equally for Klenow extension as described previously. The dsDNA library was purified and concentrated (Qiagen MinElute PCR purification kit), and further size-selected on a 20% (19:1) 1X Tris-Glycine native gel. DNA band at correct size (74-bp) was excised, electroeluted, and purified. The concentration was quantified by A260 and diluted to 1 µM in buffer EB.

The *Spec-seq* of dCas9-RNP was performed essentially as a single round of *SELEX* experiment with slight modification. Briefly, a 40 µl binding reaction containing 200 nM dCas9-RNP and 50 nM *Spec-seq* DNA library (4:1 ratio) was incubated at 37ºC for 1 hour, and resolved on a 4~20% gradient gel. Both bound and unbound DNA were excised, electroeluted, purified, and diluted to 40 µl in buffer EB. Recovered DNA from each fraction was amplified using TSSR2 and TSSR-RPIX primers to prepare sequencing library.

Sequence-specific Endonuclease Activity Measurement (SEAM-seq). *SEAM-seq* of WT-Cas9-RNP (Cas9-RNP) was performed in a 50 µl reaction containing 50 nM *Spec-seq* DNA library with either 200 nM Cas9-RNP or reaction buffer as mock treated. After 1 hour of incubation at 37 ºC, EDTA was added (60 mM final) to both digested and input samples, followed by addition of 10 µl Proteinase K (20 mg/ml, Thermo) to digest Cas9 for 30 minutes at room temperature. The remaining DNA was purified (Qiagen MinElute) and eluted in 40 µl. The full-length, uncut DNA library is PCR amplified and deep-sequenced.

Mathematical modelling of Spec-seq. Models of *Spec-seq* have been described in literature. Only a briefly outline is described here. The binding of dCas9-RNP to DNA probe i can be described using a one-site binding model:

$$DNA_i + dCas9 \overset{K_A^i}{\leftrightarrow} dCas9{:}DNA_i$$

Here, the association constant is $K_A^i = \frac{[dCas9{:}DNA_i]}{[DNA_i][dCas9]}$. We next assume that read counts in the bound and unbound *Spec-seq* libraries are:

$$n_{\text{bound},i} = A\,[dCas9{:}DNA_i]$$

$$n_{\text{unbound},i} = B\,[DNA_i]$$

Here A and B are proportional to the sequencing depth. The relative binding affinity of dCas9-RNP to sequence i relative the consensus sequence (i=1) can then be computed using:

$$K_{A,\text{rel}}^i = \frac{K_A^i}{K_A^1} = \frac{[dCas9{:}DNA_i]}{[dCas9{:}DNA_1]} \Big/ \frac{[DNA_i]}{[DNA_1]} = \frac{n_{\text{bound},i}}{n_{\text{bound},1}} \Big/ \frac{n_{\text{unbound},i}}{n_{\text{unound},1}}$$

In the simplest binding model we assumed that the relevant sequence features give linear and independent contributions to free energy of binding $\Delta\Delta G_i$:

$$K_{A,\text{rel}}^i = e^{-\frac{\Delta\Delta G_i}{RT}} = e^{\vec{X}_i \cdot \vec{\beta}_{\text{aff}}}$$

Here $\vec{X}_i = \{1, \delta_i^{A,1}, \delta_i^{C,1}, \dots\}$ is the design matrix for sequence i, $\delta_i^{n,x}$ is an indicator function for nucleotide n at position x, and the parameter vector $\vec{\beta}_{\text{aff}}$ is the PSAM. Dinucleotide interactions were included by appending the indicators $\{\delta_i^{AA,1}, \delta_i^{AC,1}, \dots\}$ to $\vec{X}_i$. Non-specific binding was modeled using the formula

$$K_{A,rel}^i = e^{-\frac{\Delta\Delta G_i}{RT}} = e^{\beta_{NS}} + e^{\vec{X}_i \cdot \vec{\beta}}.$$

The optimal models were identified by minimizing the squared $\Delta\Delta G$ error.

Mathematical modelling of SEAM-seq. The enzymatic reaction of Cas9-RNP mediated DNA digestion has previously been described using the model:

$$DNA_i{:}Cas9 \overset{k}{\rightarrow} DNA_{i,1} + DNA_{i,2}{:}Cas9$$

After the DNA target is cleaved, the Cas9-RNP remains bound to a cleavage product, resulting in a constant reduction of available Cas9-RNP during the cleavage reaction. The concentration of bound DNA

$$[DNA_i{:}Cas9] = [DNA_i]_T \frac{[Cas9]}{[Cas9] + 1/K_A^i}$$

where $[DNA_i]_T$ is the concentration of undigested DNA. To simplify the modelling of the *SEAM-seq* data, we kept the total Cas9-RNP concentration in excess of total DNA in the reaction ($[Cas9]_T \gg [DNA]_T$). Under these conditions, the free Cas9-RNP concentration is largely unchanged during the course of reaction. The rate of digestion is then:

$$\frac{d[DNA_i]_T(t)}{dt} = -k_i[DNA_i{:}\,Cas9](t) = -k_i \frac{[Cas9]}{[Cas9] + \frac{1}{K_A^i}}[DNA_i]_T(t) \equiv -k_{i,\mathrm{eff}}[DNA_i]_T(t)$$

The depletion in DNA concentration between the beginning and the end of the digestion is then found by solving this differential equation:

$$\frac{[DNA_i]_T(t_E)}{[DNA_i]_T(t_0)} = e^{-k_{i,\mathrm{eff}}t_E}$$

where t_E is the duration of the digestion. We next modeled the number of reads in the digested and input *SEAM-seq* libraries as

$$n_{\mathrm{digested},i} = C\ [DNA_i]_T(t_E)$$

$$n_{\mathrm{input},i} = D\ [DNA_i]_T(0)$$

where the constants C and D are proportional the sequencing depth. The difference in cut rate between a probe i and a reference probe ($i = 1$) can then be estimated from the data using

$$e^{(k_{i,\mathrm{eff}}-k_{1,\mathrm{eff}})(t_E-t_0)} = \frac{[DNA_i]_T(t_E)}{[DNA_i]_T(0)} \Big/ \frac{[DNA_1]_T(t_E)}{[DNA_1]_T(0)} = \frac{n_{\mathrm{digested},i}}{n_{\mathrm{input},i}} \Big/ \frac{n_{\mathrm{digested},1}}{n_{\mathrm{input},1}}$$

or, equivalently,

$$k_{i,\mathrm{eff}} - k_{1,\mathrm{eff}} = \left(\ln \frac{n_{\mathrm{digested},i}}{n_{\mathrm{input},i}} \Big/ \frac{n_{\mathrm{digested},1}}{n_{\mathrm{input},1}}\right) \Big/ t_E$$

For convenience, we define the relative nuclease activity to be the dimensionless component of the cutting rate

$$Relative\ nuclease\ activity\ (k_{app}) = \ln \frac{n_{\mathrm{digested},i}}{n_{\mathrm{input},i}} \Big/ \frac{n_{\mathrm{digested},1}}{n_{\mathrm{input},1}}$$

RESULTS

The dCas9-RNP has substantial specificity over a 23bp footprint. To determine the DNA binding specificity of the catalytically deactivated dCas9-RNP over a large footprint in a comprehensive, unbiased fashion, we performed *SELEX-seq* (**Figure 1A**). Briefly, we first assembled dCas9 bound to a gRNA (dCas9-RNP) designed against the Nanog gene (12), a gRNA that has been shown to have thousands of off-target sites. We then incubated the dCas9-RNP (0.1 μM) with a 30-bp randomized library (1 μM), flanked by PCR primer sites. The bound fraction was separated on a native gel,

recovered, amplified, and incubated with dCas9-RNP again in the same molar ratio. This isolation and enrichment was repeated for a total of five rounds after which the input and Round 5 libraries were sequenced. The clear enrichment of high-affinity sites from *SELEX-seq* Round 5 (R5, **Figure 1B)**, indicated that selection was successful.

We processed the *SELEX-seq* data using the *SelexGLM* R package (Zhang, in revision). This software package uses a biophysical model of protein binding to infer how mononucleotide features within the binding site contribute to the free energy of binding. The resulting position-specific affinity matrix (PSAM) encodes the energetic contribution of each nucleotide to affinity, and can be visualized as an energy logo (29). From this PSAM, the relative K_D (or K_A) can be calculated for any sequence by adding the differences in free energy from mismatches compared to the perfect match. *SelexGLM* can model the binding affinity over a large footprint, which is essential for understanding dCas9-RNP binding.

The energy logo for the dCas9-RNP has two striking features (**Figure 1C**): 1) the expected preference of the dCas9-RNP for an NGG PAM site; and 2) the lack of sequence preference in the PAM distal region (13-20) of the protospacer. There is a clearly defined region of sequence preference within the PAM-proximal protospacer that is consistent with a seed sequence confined to the PAM-proximal driving binding. Most specificity is contained in PAM-proximal bases 1-11, with mild preference for matches at positions 12-13. Within the PAM-proximal region, the greatest binding energy is derived from exact matches, with some tolerance at positions 6 and 7 for a G and T mismatch, respectively. The dCas9-RNP also exhibits some preference for Gs downstream of the PAM (–5 to –8) and Ts further downstream (–10 to –12). Although minor compared to the affinity contributed by the PAM and PAM-proximal base pairs (~6% *vs.* 92% of total specificity computed by maximal ΔΔG/RT), these downstream sequences contribute more to DNA binding affinity than PAM-distal base pairs (0.5 kcal/mol vs. 0.2 kcal/mol). Despite the lack of preference in the PAM-distal region, the dCas9-RNP nonetheless is sensitive to base identity over a 23-bp footprint.

The PSAM (**Figure 1C**) not only reveals which bases are preferred by the dCas9-RNP, but also the mismatches that are most disfavored. The gRNA used in this experiment is highly purine-rich within the protospacer (15/20 positions, including 12/3 within the PAM-proximal region). At these purine positions, the most detrimental mismatches are non-complementary pyrimidines (A->C, G->T). These mismatches, for example C instead of A, would result in the A of the gRNA attempting to base pair with a G on the opposite strand, which is both non-complementary and bulky. This pattern is remarkably consistent, and indicates that dG·rA and dA·rG base pairing between the gRNA and protospacer is the most detrimental to binding. An exception to this is the T at position 6, which also tolerates a C least. This C would also result in a non-complementary G opposite the T in the gRNA. Together, it is clear from these data that not all mismatches have an equal effect on binding, with some more detrimental than others.

<u>*Parallel measurement of DNA binding and cleavage using Spec-seq and SEAM-seq.*</u> Cas9-RNP activity involves the two-step process of binding followed by cutting DNA (**Figure 2A)**. The overall efficiency of the enzyme depends on how well it binds a sequence (K_A) then how quickly it is cut (turnover or k_{cat}).

To generate these two values, we paired the techniques of *Spec-seq* (24, 25) to measure affinity and *SEAM-seq* (this work) to measure endonuclease activity (**Figure 2B,C**, described in detail below). Performing these assays under identical conditions on a single library, composed of seven sub-libraries encompassing a total of 1,594 sequences, allowed direct comparison of the effect of mismatches on binding and nuclease activity for each sequence. By fitting both data sets to sequence features, specificity models for both binding and cleavage can be generated. Further, we are able to estimate the relative occupancy of sequences based on the measured affinity in *Spec-seq*, calculate the turnover of the enzyme, and generate a general model for the enzyme efficiency throughout the protospacer, PAM, and flanking regions. This general efficiency model can then be used to estimate off-target cutting at sites that differ from the intended target.

The SELEX-seq and Spec-seq binding specificity measurements are consistent. To measure the relative affinity of the dCas9-RNP for each sequence, we performed standard *Spec-seq* (24, 25) using the dCas9-RNP. Sub-libraries were pooled equally and incubated with dCas9-RNP at a 1:4 ratio (50 nM DNA: 200 nM RNP). Bound and unbound fractions were then separated on a native gel (**Figure S1A**), recovered, and sequenced. After ensuring that sub-libraries are similarly represented in the full library (**Figure S2C**), the relative binding affinity of dCas9-RNP for each site were calculated directly by taking the ratio of the probe frequencies in the bound and unbound fractions (See Methods). At a sequencing depth of ~30 million reads per sample, each sequence was represented by an average of ~20,000 counts. This allowed quantification of affinities across four orders of magnitude, as compared to ~2 orders of magnitude for *SELEX-seq* followed by *SelexGLM* analysis (**Figure 3C, E, Figure S1C**).

The results obtained using *Spec-seq* agreed very well with the PSAM obtained using *SelexGLM*. In particular, we observed a high correlation between the directly measured *Spec-seq* sequences compared to the calculated affinity from the *SelexGLM* PSAM (Spearman $\rho = 0.88$, $R^2 = 0.85$, **Figure S1C**). Inconsistencies between the two methods are likely due to two main factors: sequencing depth and any non-additive effect of multiple mismatches on affinity. Capturing the ~10,000-fold difference in affinity measured by *Spec-seq* requires differences in reads for *SELEX-seq* scaled to the round of selection (R5 in this case), or a sequencing depth of 10^{20}. As a result, low-affinity sequences are not captured by *SelexGLM*, resulting in apparent compression compared to *Spec-seq*. Also, because *SelexGLM* assumes that all positions are independent, it likely underestimates the effect of multiple mismatches, whereas these effects are directly measured by *Spec-seq*. Despite these differences, a linear regression model fit to the *Spec-seq* data **(Figure S4A, S6B)** and the *SelexGLM* model **(Figure 1C, S6A)** are remarkably consistent. Thus, like the *SelexGLM* model built from *SELEX-seq* data, *Spec-seq* analysis confirms that purine:purine mismatches are the least favorable for binding, the PAM-distal region does not contribute to affinity, and that the closer a mismatch is to the PAM, the more deleterious the effect is on binding (**Figure 3E**). However, there are also exceptions. For example, some PAM-proximal mismatches have very little effect on affinity (Libraries 1, 2, and 3, **Figure 3E**), such as the change of guanine at position of 4 of protospacer to adenine (G4A), T6A, and G9A ($\ln K_A$: -0.01±0.27, 0.08±0.26, -0.12± 0.06 respectively). These examples are reproducible, indicating that some PAM-proximal region mismatches can be tolerated by the dCas9-RNP.

Sequence-specific endonuclease activity measured by sequencing (SEAM-seq). To accurately measure the relative endonuclease activity of WT-Cas9-RNP (Cas9-RNP), we developed *SEAM-seq*, a method explicitly designed to be paired with *Spec-seq* (24, 25) (**Figure 2B**). First, the same *Spec-seq* library (50 nM) was used to measure the binding specificity of the dCas9-RNP is incubated with a 4x excess of Cas9-RNP (200 nM), and allowed to cleave for 1 hour under the same conditions that we ran *Spec-seq*. At this nuclease:DNA ratio, high affinity sites (K_D~1 nM to 100 nM) are expected to be fully occupied by the enzyme, whereas lower affinity sites (K_D > ~100 nM) only partially occupied. At the end of incubation, the remaining uncut DNA was amplified by PCR. A control mock digestion (input) was performed by incubating library with buffer, and subjected to the same downstream processes. Both the uncut and the input libraries were then sequenced. With each library being similarly represented (**Figure S2**), the ratio of reads between uncut and input libraries measures the relative fraction cut, or relative depletion of each DNA sequence by the Cas9-RNP. The logarithm of this depletion ratio is proportional to the difference of apparent relative rate of cleavage (or Δk_{app}) (**Figure 2B, C;** see Method section for details). To establish a reference baseline for the relative cut rate, we set the average cut rate of all exact matches within the PAM and protospacer to zero, and calculated the relative impact of mismatches on the endonuclease activity. Measured in this way, sequences with a $k_{app} < 0$ are cleaved less than the perfect match, whereas those with a $k_{app} > 0$ are cleaved more during the 1 hour incubation. Two experimental replicates of *SEAM-seq* showed that the technique is highly reproducible (**Figure S1E**, Spearman's $\rho = 0.93$, $R^2 = 0.96$), with the average k_{app} used for downstream analyses.

Base pairing throughout the protospacer is important for endonuclease activity. To investigate how mismatches disrupt endonuclease activity, we modeled the endonuclease activity using linear regression with mononucleotides features as predictors (**Figure S5A**). This revealed that endonuclease activity requires base pairing over almost the entire protospacer (mismatches at position 20 have a less prominent effect) (**Figure 3D, F, S5A**). This is consistent with the previous model where base pairing in the PAM-distal region is critical for endonuclease activity (7), and in marked contrast to the more localized effect of each nucleotide on binding affinity (1). Analogous to their effect on binding affinity, the sequence logo revealed that purine:purine mismatches tend to disrupt endonuclease activity most strongly. Lastly, although sequences downstream of the PAM contribute to affinity, variation in this region does not affect endonuclease activity significantly.

Most combinations of mismatches in the PAM-proximal region decrease affinity and activity. Plotting the binding affinity measured by *Spec-seq* against the endonuclease activity measured by *SEAM-seq* showed that most (871 out of 1007) sequences with reduced binding also exhibited reduced nuclease activity (**Figure 4A**). This correlation is, as expected, strongest for mismatches within the PAM and the PAM-proximal region (positions -3 to +11, black, red, blue, and part of green libraries) where mismatches result in lower affinity (**Figure 4B-E**). This correlation between affinity and activity is most striking when the affinity is ~150x less ($\ln K_A < -5$) than the perfect complement. At the concentration used in the experiment, and the K_D of the Cas9-RNPs for the perfect complement is ~1 nM, the reduced endonuclease activity over those low affinity sites is primarily due to reduced occupancy.

Affinity does not always correlate with endonuclease activity. Three classes of mismatches do not have corresponding reductions in affinity and activity: 1) those at PAM-distal positions (+12 to +20); 2) those downstream of the PAM (-4 to -8); and 3) single and multiple mismatches in the PAM-proximal region that enhance endonuclease activity. First, consistent with previous observations (7), mismatches in the PAM-distal region (12-20, yellow, grey and part of green libraries) have little effect on binding affinity, but universally impair endonuclease activity (**Figure 3F**). Second, specific sequences downstream of the PAM enhance affinity but do not affect activity. These sequences, if present in the genome, would have enhanced endonuclease activity, but can be easily identified and avoided.

Most interesting is the third class, which consist of mismatches in the PAM-proximal region that enhance endonuclease activity, a phenomenon we term mismatch-activation. These sequences can be most easily visualized in the breakout plots for each individual library (black, red, blue, part of green, positions -3 to 11, and **Figure 4B-G**). As discussed above, most mismatches within this region (649 out of 781) impair both binding and endonuclease activity and fall along a binding sigmoid from the origin to the lower left of the plots. However, several sequences lie above this sigmoid, indicating enhanced endonuclease activity. Further, some of these sequences are cut more completely than even the perfect match ($k_{app} > 0$), even while affinity is lower (relative $\ln K_A < 0$). A group of these mismatch-activated sequences is particularly evident in the blue library (positions 6 to 9). Within this library, the activity against some sequences are ~1.5x greater than the perfect match, indicating that sequences that don't match the gRNA can strongly enhance enzyme activity (**Figure S2A**). Among these mismatch-activated sequences, 26 of 27 have one or two mismatches, with the remaining sequence having three mismatches (**Figure S2B**). Under the conditions of the experiment, the mismatch-activation sequences are likely to be fully occupied by the Cas9-RNP, indicating that the increased activity is due to increased turnover and not affinity.

Single mismatches in the PAM proximal region almost universally enhance Cas9-RNP endonuclease activity (**Figure 4I**). This single base pair mismatch-activation reaches a peak with mismatches between positions 3 and 9 (**Figure 4I**), but remains significant at other positions in the region. Though activating, these single mismatches nonetheless reduce affinity (**Figure 4H**). By contrast, pairs of adjacent mismatches throughout this region reduce affinity synergistically (**Figure 4J**), whereas the same pairs have a less that additive effect on mismatch-activation (**Figure 4K**). These findings suggest that destabilizing the RNA:DNA helix can enhance enzymatic activity even while reducing affinity. When the helix is destabilized further by adjacent mismatches, enhancement is still observed, but the penalty for binding continues to increase, reducing the probability of off-target cutting.

The addition of DNA features improves the Spec-seq affinity model. Although *Spec-seq* and *SEAM-seq* directly measure the activity of ~1,600 sequences, few of these are found in the genome. We therefore sought to develop general models that could predict off-target binding and cutting for sequences not present in our library. To predict the binding affinity, we first assumed that the binding free energy ΔΔG is due to independent linear contributions from each nucleotide in the binding site (described above). This generated a PSAM (**Figure 5A**) that agreed well with *SELEX-seq* PSAM and the model predicted the measured ΔΔG values quite well (**Figure S4A**, model *Mono*, $R^2 = 0.95$). However, plotting predicted

versus measured ΔΔG values revealed significant model bias; the affinity of low-affinity probes was underestimated and the range of predicted valued for higher-affinity probes was compressed (**Figure S4A**). We attempted to improve our model by adding a non-specific binding component that moderates the predicted affinity of poorly matched sequences. This dramatically improved the fit and removed most of the bias (**Figure S4B**, model *Mono/NS*, R^2 = 0.98). As noted above, we observed a non-linear contribution of adjacent dinucleotide mismatches to binding (**Figure 4J**). When we incorporated the dinucleotide interactions into the model, the fit improved still further (**Figure 5A, S4C**, model *M/Di/NS*, R^2 = 0.99). This method not only generates a very accurate model for the effect of any mismatch on binding but also validates the finding that adjacent mismatches have a greater than additive effect on affinity (8).

Calculating the catalytic activity of Cas9 by integrating Spec-seq and SEAM-seq. To generate a general model for apparent endonuclease activity (k_{app}) we started with a linear model that used only mononucleotide features. This simple model described the *SEAM-seq* data reasonably well but also exhibited significant bias (**Figure S5A**, model *Mono*, R^2 = 0.77), suggesting that a more refined model of k_{app} was needed. The k_{app} measured by *SEAM-seq* reflects a two-step process (**Figure 2A**): binding (K_A) followed by affinity-independent cutting (k_{cat}). Because the binding of higher affinity sequences was more saturated than lower affinity sequences, we reasoned that correcting for occupancy would enable estimation of k_{cat}. Having run the *SEAM-seq* and *Spec-seq* experiments under matched experimental conditions, we were able to determine the relative occupancy of Cas9 by assuming that the occupancy has a sigmoidal dependence on the binding affinity and included the free protein concentration as a free parameter. This allowed calculation of model for k_{cat}, and allowed generation of a model for Cas9-RNP efficiency for off-target sequences.

The model for k_{cat} (**Figure 5C**), which represents affinity-independent cleavage, significantly improved our modeling of observed k_{app}, when used in combination with the relative occupancy predicted by affinity model (**Figure S5E, bottom plot**). As expected, because mismatches do not affect affinity in the PAM-distal region, the sequence logo of k_{cat} shows a negative effect on activity for any mismatches in the PAM-distal region. Conversely, purine:purine mismatches from positions 6-9 within the PAM-proximal region have increased activity, mirroring the mismatch-activation described above. Surprisingly, the mismatches with the greatest activation (C and T) exhibit the most negative effect on affinity. Thus, although mismatches can enhance activity of the Cas9-RNP, increased off-target cleavage is offset by decreased binding to some extent. Finally, the improved performance of the model suggests that it can be used to predict the k_{app} for the Cas9-RNP at different protein concentrations.

DISCUSSION

As Cas9 and other RNA programmed endonucleases are increasingly widely used in genomic experiments, there is a pressing need to be able to predict off-target activity. The specificity of an enzyme is a function of the efficiency for off-target sites and the concentration of enzyme and ligand. Thus, predicting off-target cutting requires a measurement of efficiency, or the product for the turnover rate of the enzyme (k_{cat}) and the propensity to bind (estimated by K_A). Previous experiments have

identified sites that can be cut in the genome or from random libraries (17, 19, 30, 31), and have also identified alternate binding sites from genomic experiments (12, 13). Although these have provided information about where the Cas9-RNP can cut, no clear rule can be summarized from any of these assays to predict where the Cas9-RNP will cut. Here we present a systematic experimental and computational pipeline to measure Cas9-RNP efficiency for on- and off-target sites that can be performed rapidly for any gRNA.

To accurately measure the specificity of a Cas9-RNP, we produced the first free-energy based DNA binding model. Using a gRNA directed at the Nanog gene, which is known to have thousands off-target binding sites (12), we used our *SELEX-seq/SelexGLM* pipeline (Zhang, in revision, (23) to calculate a comprehensive, unbiased PSAM model for binding any sequence. This model, from which the relative K_A of any sequence can be calculated, confirmed and provided more details than previous models, reinforcing the idea that the PAM, and PAM-proximal sequences (positions 1-13) are the determinants of binding affinity. The model also revealed a region downstream of the PAM that contributes to specificity in sequence-dependent fashion. Finally, the PSAM also revealed, as had been suggested previously (16, 32), that not all mismatches have the same effect on affinity, with purine:purine mismatches between gRNA and targeted strand of DNA being the most deleterious. Importantly, this model was recapitulated using a much simpler and less labor-intensive technique, *Spec-seq*, that can be paired with our endonuclease activity assay, *SEAM-seq*.

In addition to thoroughly validating the *SelexGLM* PSAM model, the *Spec-seq* results provided insight into the mechanism of binding. Due to the depth of sequencing, we were able to use linear regression to incorporate the contribution of non-specific binding and dinucleotide pairs to binding, which revealed that the thermodynamic contribution of each base pair to binding is not independent (**Figure 4J, 5C**). This finding and the diminished contribution of complementarity to affinity as distance from the PAM increases from positions 10-13 support the model that PAM-melting is followed by directional sequential base-pairing over ~13 sites to achieve full affinity (8). Overall, these models are a significant improvement over previous *ChIP-seq* based methods (**Figure S6C**), and can be used to estimate binding at off-target sites, even those that contain mutations (12). More importantly, the binding specificity of the dCas9-RNP suggests that a re-evaluation of how gRNAs are designed for dCas9-based genomic experiments is needed.

Binding to DNA is only the first of two steps required for DNA editing by Cas9-RNP. By pairing the relative K_A values with apparent endonuclease activity (k_{app}) using *SEAM-seq*, it is now possible to decouple the effects of mismatches on affinity and activity. Although in most cases the two steps are correlated (e.g. decreased affinity leads to decreased cutting), there are systematic exceptions that shed light on the mechanism. First, as has been observed previously, it is clear that mismatches in the PAM-distal region impair enzymatic activity while affinity is unaffected. Interestingly, multiple mismatches have a less than additive effect on activity, indicating that an initial break in base pairing is the key step (**Figure 4K**). Second, and perhaps most strikingly, mismatches within the PAM-proximal region of protospacer, while they decrease affinity, actually enhance enzymatic activity. Importantly, adjacent mismatches within this region decrease binding synergistically, while enhancing activity less

than additively. This indicates that the loss in binding affinity of mismatches can be partially compensated by enhanced activity. This has implications in designing gRNAs to cut specifically in the genome, but also reveals that the Cas9 system for bacterial defense against viral invasion has a built-in tolerance for mismatches in the PAM-proximal region of the protospacer, possibly due to an evolutionary pressure to adapt to mutations in viral genomes. These findings, that binding does not always result in efficient cutting, help to explain the lack of overlap between where the Cas9-RNP binds the genome, and where it cuts (12, 13, 30).

Mismatch activation is particularly prominent between positions 6 and 9 of the PAM-proximal region. Structural studies have shown that the DNA:RNA duplex takes a sharp bend in this region (**Figure 3B**). Mismatch-activation has been noted in previous experiments using other two different gRNAs (33). Although not observed as consistently, this phenomenon was nonetheless most prominent in same region (positions ~5-10) (33). However, it was not known at that time whether the increased cutting was due to increased enzyme turnover or enhanced binding. Whether mismatches enhance activity by facilitating this bend awaits further study.

Finally, with a pipeline to measure affinity and activity under the same conditions in hand, the relative efficiency of a Cas9-RNP can now be estimated using regression-based modeling of *Spec-seq* and *SEAM-seq* data. By correcting for baseline activity and occupancy, k_{cat} can be calculated, enzymatic efficiency can be calculated as the product of K_A and k_{cat}. Based on this model, it is clear that mismatches in the PAM-distal region create an enzyme that is less efficient. It is also clear that, although mismatches within the PAM-proximal region enhance activity, they cause similar overall decreases in efficiency. As a result, any single mismatch in the PAM-distal and late (positions 6-9) PAM-proximal region can have similar effects on efficiency. More PAM-proximal mismatches are even less efficient. Thus, in designing gRNAs, those with other genomic sites containing mismatches in the near PAM-proximal region will have a lower probability for off-target cutting than those in the more PAM-distal regions, which still retain significant efficiency.

In addition to providing insight into how mismatches affect the mechanism of Cas9-RNP binding and cleavage, we describe a simple experimental protocol and pipeline for determining the specificity of Cas9:gRNA pair. Because the *SELEX-seq* generated PSAM is recapitulated so well by the *Spec-seq* model for binding, it is reasonable to design libraries similar to those described here based on gRNA sequence alone. Compared to other high-throughput measurements of binding (such as: CHAMP(34) or Hits-Flip(35)), *Spec-seq* and *SEAM-seq* require no special equipment, and multiple gRNAs can be assessed in parallel as described with sequencing-ready libraries in a few days. The computational demands are also minimal, and the scripts provided can be run on the resulting data, with results in a few hours. The robust models generated by this technique will allow estimation of off-target binding and cleavage for any gRNA, modified Cas9, or other sequence-specific endonuclease to guide the design of the most appropriate tool for specific applications.

ACCESSION NUMBERS

Sequencing reads will be deposited prior to publication.

ACKNOWLEDGEMENT

We thank members of the Pufall and Bussemaker labs for useful discussions. We also thank Dipa Sashital and Stephen Floor for critically reading the manuscript and providing helpful input.

FUNDING

This research was supported by the National Science Foundation CAREER grant 1552862 (M.A.P), and grant R01HG003008 (H.J.B) from National Institutes of Health.

CONFLICT OF INTEREST

The authors have no competing financial of personal interests to declare.

SUPPLEMENTARY DATA

Supplementary figures and data can be found in the attached Supplementary document.

REFERENCES

1. Jinek,M., Chylinski,K., Fonfara,I., Hauer,M., Doudna,J.A. and Charpentier,E. (2012) A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. *Science*, **337**, 816–821.

2. Cong,L., Ran,F.A., Cox,D., Lin,S., Barretto,R., Habib,N., Hsu,P.D., Wu,X., Jiang,W., Marraffini,L.A., *et al.* (2013) Multiplex genome engineering using CRISPR/Cas systems. *Science*, **339**, 819–823.

3. Chen,B., Gilbert,L.A., Cimini,B.A., Schnitzbauer,J., Zhang,W., Li,G.-W., Park,J., Blackburn,E.H., Weissman,J.S., Qi,L.S., *et al.* (2013) Dynamic imaging of genomic loci in living human cells by an optimized CRISPR/Cas system. *Cell*, **155**, 1479–1491.

4. Hilton,I.B., D'Ippolito,A.M., Vockley,C.M., Thakore,P.I., Crawford,G.E., Reddy,T.E. and Gersbach,C.A. (2015) Epigenome editing by a CRISPR-Cas9-based acetyltransferase activates genes from promoters and enhancers. *Nat. Biotechnol.*, **33**, 510–517.

5. Konermann,S., Brigham,M.D., Trevino,A.E., Joung,J., Abudayyeh,O.O., Barcena,C., Hsu,P.D., Habib,N., Gootenberg,J.S., Nishimasu,H., *et al.* (2015) Genome-scale transcriptional activation by an engineered CRISPR-Cas9 complex. *Nature*, **517**, 583–588.

6. Qi,L.S., Larson,M.H., Gilbert,L.A., Doudna,J.A., Weissman,J.S., Arkin,A.P. and Lim,W.A. (2013) Repurposing CRISPR as an RNA-guided platform for sequence-specific control of gene expression. *Cell*, **152**, 1173–1183.

7. Jinek,M., Chylinski,K., Fonfara,I., Hauer,M., Doudna,J.A. and Charpentier,E. (2012) A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. *Science*, **337**, 816–821.

8. Szczelkun,M.D., Tikhomirova,M.S., Sinkunas,T., Gasiunas,G., Karvelis,T., Pschera,P., Siksnys,V. and Seidel,R. (2014) Direct observation of R-loop formation by single RNA-guided Cas9 and Cascade effector complexes. *Proceedings of the National Academy of Sciences*, **111**, 9798–9803.

9. Sternberg,S.H., Redding,S., Jinek,M., Greene,E.C. and Doudna,J.A. (2014) DNA interrogation by the CRISPR RNA-guided endonuclease Cas9. *Nature*, **507**, 62–67.

10. Sternberg,S.H., LaFrance,B., Kaplan,M. and Doudna,J.A. (2015) Conformational control of DNA target cleavage by CRISPR–Cas9. *Nature*, **527**, 110–113.

11. Jiang,F. and Doudna,J.A. (2017) CRISPR–Cas9 Structures and Mechanisms. *Annu. Rev. Biophys.*, **46**, 505–529.

12. Wu,X., Scott,D.A., Kriz,A.J., Chiu,A.C., Hsu,P.D., Dadon,D.B., Cheng,A.W., Trevino,A.E., Konermann,S., Chen,S., *et al.* (2014) Genome-wide binding of the CRISPR endonuclease Cas9 in mammalian cells. *Nat. Biotechnol.*, **32**, 670–676.

13. Kuscu,C., Arslan,S., Singh,R., Thorpe,J. and Adli,M. (2014) Genome-wide analysis reveals characteristics of off-target sites bound by the Cas9 endonuclease. *Nat. Biotechnol.*, **32**, 677–683.

14. Boyle,E.A., Andreasson,J.O.L., Chircus,L.M., Sternberg,S.H., Wu,M.J., Guegler,C.K., Doudna,J.A. and Greenleaf,W.J. (2017) High-throughput biochemical profiling reveals sequence determinants of dCas9 off-target binding and unbinding. *Proceedings of the National Academy of Sciences*, **114**, 5461–5466.

15. Fu,Y., Foden,J.A., Khayter,C., Maeder,M.L., Reyon,D., Joung,J.K. and Sander,J.D. (2013) High-frequency off-target mutagenesis induced by CRISPR-Cas nucleases in human cells. *Nat. Biotechnol.*, **31**, 822–826.

16. Hsu,P.D., Scott,D.A., Weinstein,J.A., Ran,F.A., Konermann,S., Agarwala,V., Li,Y., Fine,E.J., Wu,X., Shalem,O., *et al.* (2013) DNA targeting specificity of RNA-guided Cas9 nucleases. *Nat. Biotechnol.*, **31**, 827–832.

17. Kim,D., Bae,S., Park,J., Kim,E., Kim,S., Yu,H.R., Hwang,J., Kim,J.-I. and Kim,J.-S. (2015) Digenome-seq: genome-wide profiling of CRISPR-Cas9 off-target effects in human cells. *Nat Meth*, **12**, 237–243.

18. Schaefer,K.A., Wu,W.-H., Colgan,D.F., Tsang,S.H., Bassuk,A.G. and Mahajan,V.B. (2017) Unexpected mutations after CRISPR-Cas9 editing in vivo. *Nat Meth*, **14**, 547–548.

19. Cameron,P., Fuller,C.K., Donohoue,P.D., Jones,B.N., Thompson,M.S., Carter,M.M., Gradia,S., Vidal,B., Garner,E., Slorach,E.M., *et al.* (2017) Mapping the genomic landscape of CRISPR-Cas9 cleavage. *Nat Meth*, 10.1038/nmeth.4284.

20. Fu,Y., Sander,J.D., Reyon,D., Cascio,V.M. and Joung,J.K. (2014) Improving CRISPR-Cas nuclease specificity using truncated guide RNAs. *Nat. Biotechnol.*, **32**, 279–284.

21. Doench,J.G., Hartenian,E., Graham,D.B., Tothova,Z., Hegde,M., Smith,I., Sullender,M., Ebert,B.L., Xavier,R.J. and Root,D.E. (2014) Rational design of highly active sgRNAs for CRISPR-Cas9-mediated gene inactivation. *Nat. Biotechnol.*, **32**, 1262–1267.

22. Zheng,T., Hou,Y., Zhang,P., Zhang,Z., Xu,Y., Zhang,L., Niu,L., Yang,Y., Liang,D., Yi,F., *et al.* (2017) Profiling single-guide RNA specificity reveals a mismatch sensitive core sequence. *Sci. Rep.*, **7**, 40638.

23. Kribelbauer,J.F., Laptenko,O., Chen,S., Martini,G.D., Freed-Pastor,W.A., Prives,C., Mann,R.S. and Bussemaker,H.J. (2017) Quantitative Analysis of the DNA Methylation Sensitivity of Transcription Factor Complexes. *CellReports*, **19**, 2383–2395.

24. Stormo,G.D., Zuo,Z. and Chang,Y.K. (2015) Spec-seq: determining protein-DNA-binding specificity by sequencing. *Brief Funct Genomics*, **14**, 30–38.

25. Zuo,Z. and Stormo,G.D. (2014) High-resolution specificity from DNA sequencing highlights alternative modes of Lac repressor binding. *Genetics*, 10.1534/genetics.114.170100/-/DC1.

26. Anders,C. and Jinek,M. (2014) In vitro enzymology of Cas9. *Meth. Enzymol.*, **546**, 1–20.

27. Lin,S., Staahl,B.T., Alla,R.K. and Doudna,J.A. (2014) Enhanced homology-directed human genome engineering by controlled timing of CRISPR/Cas9 delivery. *eLife*, **3**, e04766.

28. Roven,C. and Bussemaker,H.J. (2003) REDUCE: An online tool for inferring cis-regulatory elements and transcriptional module activities from microarray data. *Nucleic Acids Research*, **31**, 3487–3490.

29. Foat,B.C., Morozov,A.V. and Bussemaker,H.J. (2006) Statistical mechanical modeling of genome-wide transcription factor occupancy data by MatrixREDUCE. *Bioinformatics*, **22**, e141–9.

30. Tsai,S.Q., Zheng,Z., Nguyen,N.T., Liebers,M., Topkar,V.V., Thapar,V., Wyvekens,N., Khayter,C., Iafrate,A.J., Le,L.P., *et al.* (2014) GUIDE-seq enables genome-wide profiling of off-target cleavage by CRISPR-Cas nucleases. *Nature Biotechnology*, **33**, 187–197.

31. Tsai,S.Q., Nguyen,N.T., Malagon-Lopez,J., Topkar,V.V., Aryee,M.J. and Joung,J.K. (2017) CIRCLE-seq: a highly sensitive in vitro screen for genome-wide CRISPR–Cas9 nuclease off-targets. *Nat Meth*, **14**, 607–614.

32. Doench,J.G., Fusi,N., Sullender,M., Hegde,M., Vaimberg,E.W., Donovan,K.F., Smith,I., Tothova,Z., Wilen,C., Orchard,R., *et al.* (2016) Optimized sgRNA design to maximize activity and minimize off-target effects of CRISPR-Cas9. *Nat. Biotechnol.*, **34**, 184–191.

33. Hua Fu,B.X., Hansen,L.L., Artiles,K.L., Nonet,M.L. and Fire,A.Z. (2014) Landscape of target:guide homology effects on Cas9-mediated cleavage. *Nucleic Acids Research*, **42**, 13778–13787.

34. Jung,C., Hawkins,J.A., Jones,S.K., Xiao,Y., Rybarski,J.R., Dillard,K.E., Hussmann,J., Saifuddin,F.A., Savran,C.A., Ellington,A.D., *et al.* (2017) Massively Parallel Biophysical Analysis of CRISPR-Cas Complexes on Next Generation Sequencing Chips. *Cell*, **170**, 35–47.e13.

35. Nutiu,R., Friedman,R.C., Luo,S., Khrebtukova,I., Silva,D., Li,R., Zhang,L., Schroth,G.P. and Burge,C.B. (2011) Direct measurement of DNA affinity landscapes on a high-throughput sequencing instrument. *Nature Biotechnology*, **29**, 659–664.

TABLE AND FIGURES LEGENDS

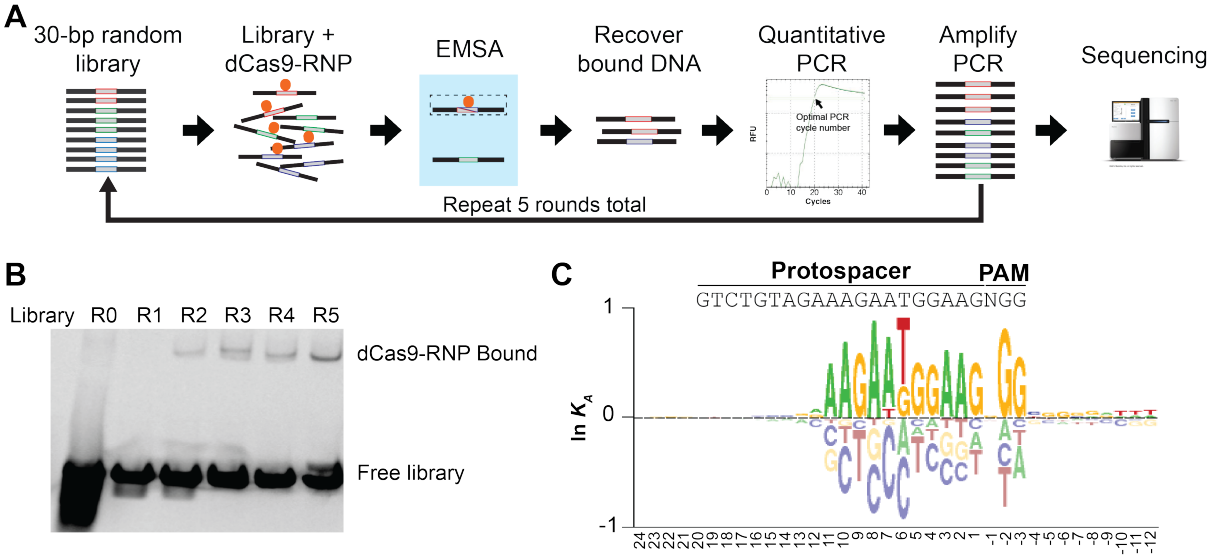
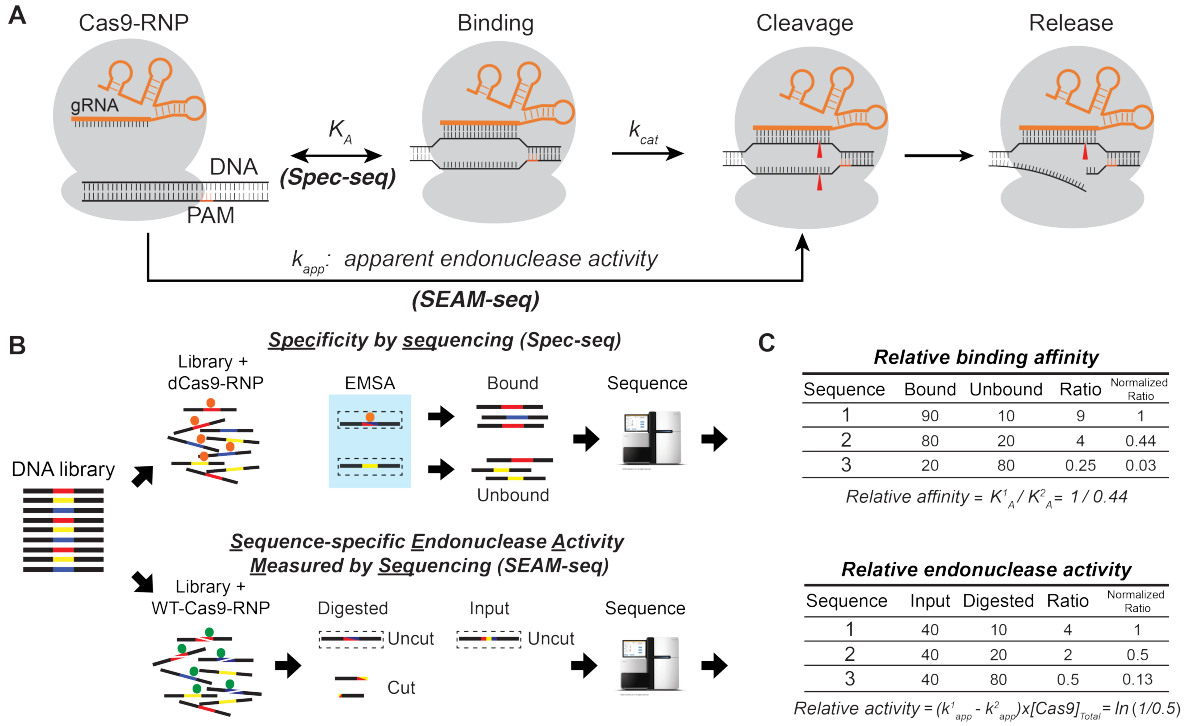


Figure 1: The DNA binding specificity of the dCas9-RNP extends over 23bp. (A) Schematic workflow of *SELEX-seq*. A 30bp random library flanked by sequencing adapters is incubated with the Cas9-RNP. Bound sequences are separated on an EMSA gel, excised, isolated, carefully amplified, and then used in the next round of selection. When performed under limiting Cas9-RNP concentrations, the round-by-round amplification is proportional to the affinity for a given sequence. **(B)** Summary EMSA gel. Equal concentrations of DNA library and Cas9-RNP are used in each lane. The increase in band intensity with each round indicates that the pool is being enriched for specific sequences. **(C)** Enrichment of sequences in the final library (R5) is compared to the initial library (R0) using a generalized linear model (*SelexGLM*) to derive a Position Specific Affinity Matrix (PSAM), reflecting the contribution of each base pair to the affinity of the Cas9-RNP for DNA. The majority of specificity is derived from the PAM (-1 to -3) and the PAM-proximal protospacer (1-11), although significant affinity is derived from PAM-distal base paring (positions 12,13) and sequence downstream of the PAM (-4 to -12).



Sequence	Bound	Unbound	Ratio	Normalized Ratio
1	90	10	9	1
2	80	20	4	0.44
3	20	80	0.25	0.03

Relative affinity = K^1_A / K^2_A = 1 / 0.44

Sequence	Input	Digested	Ratio	Normalized Ratio
1	40	10	4	1
2	40	20	2	0.5
3	40	80	0.5	0.13

Relative activity = $(k^1_{app} - k^2_{app}) x [Cas9]_{Total} = ln(1/0.5)$

Figure 2. Decoupling DNA binding and endonuclease activity of the Cas9-RNP by *Spec-seq* and *SEAM-seq*. (A) Two-step enzymatic model of Cas9-RNP-mediated DNA cleavage. The apparent endonuclease activity of the Cas9-RNP (k_{app}) is determined using *SEAM-seq*, whereas the relative affinity of binding (K_a) is measured by *Spec-seq* for each of the 1,594 sequences in the library. **(B,C)** Schematic of the *Spec-seq* and *SEAM-seq* protocols. For *Spec-seq*, sequences bound by the catalytic-deactivated Cas9-RNP are resolved from unbound on an EMSA gel. Each band is excised and deep-sequenced, with the relative affinity directly calculated **(C)** from the ratio of bound to unbound (see Materials and Methods). For *SEAM-seq*, the same sequences are cleaved by wild-type Cas9-RNP. Uncut sequences are amplified by PCR and deep-sequenced. The ratio of uncut sequences in digested to input fraction reflects the relative cleavage of sequences **(C)**, which can be normalized to calculate the relative apparent endonuclease activity (Δk_{app}, see Materials and Methods).

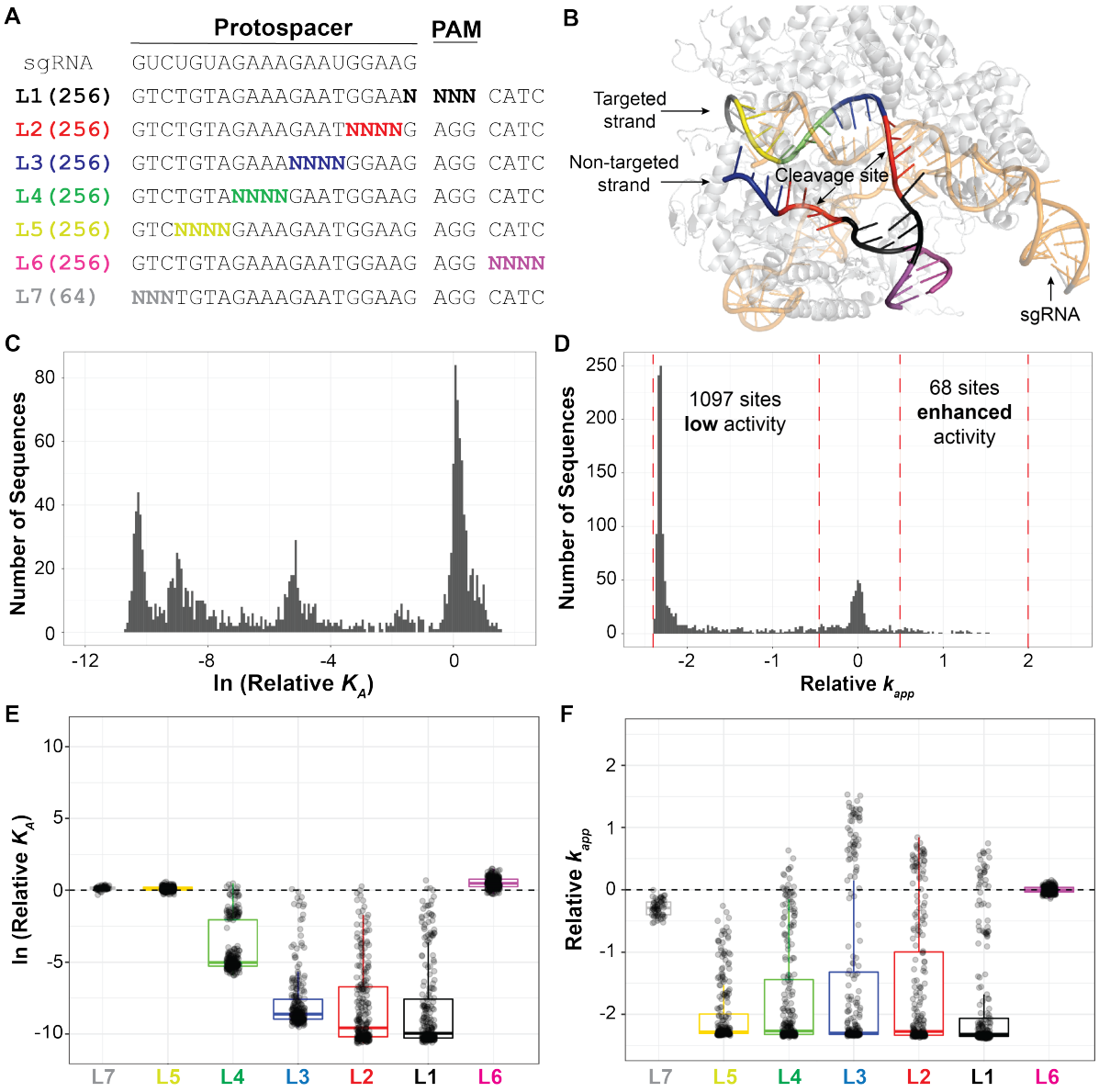


Figure 3. The sequence preference for endonuclease activity and affinity is similar, except in the PAM-distal region. (A) Seven sub-libraries were pooled to comprise the library used in *SEAM-seq* and *Spec-seq*. The sequences cover all possible single mismatches and blocks of up to 4 mismatches through and downstream of the PAM as well as the protospacer. **(B)** DNA libraries mapped on the existing crystal structure of Cas9-RNP:DNA complex (PDB: 5F9R). L3 is positioned at the bend in RNA:DNA duplex. **(C)** Histogram showing the distribution of relative binding affinity indicating that many mismatches to not cause a large effect on binding. **(D)** Histogram of the relative apparent endonuclease activity indicating that most mismatches decrease activity, although some do increase cutting (farther right). **(E, F)** Boxplots showing the distribution of the effect of mismatches on affinity and endonuclease activity. **(E)** Mismatches closer to the PAM have a more deleterious effect than distal mismatches on affinity. **(F)** Many mismatches increase endonuclease activity, particularly in L3 (positions 6-9).

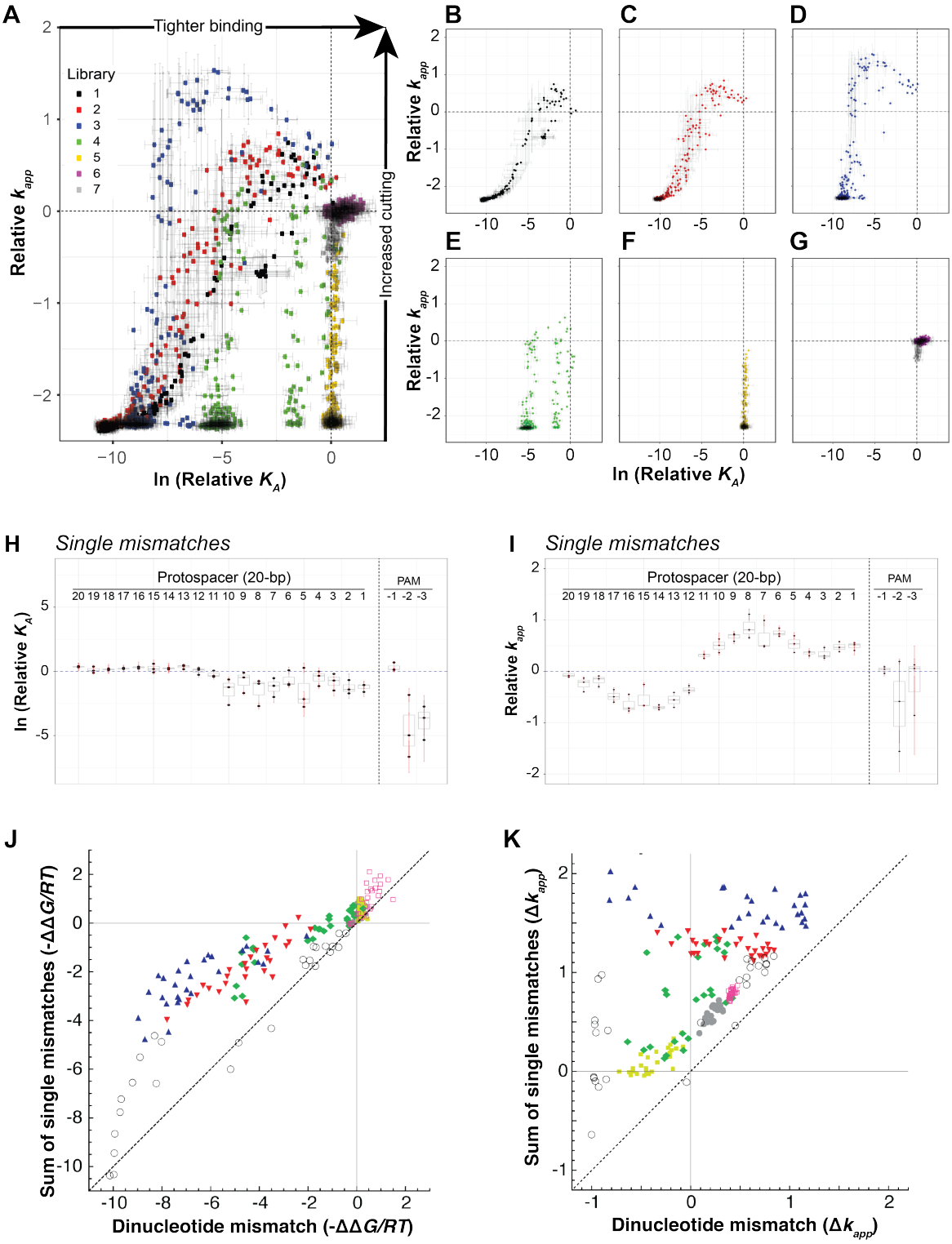


Figure 4: Dozens of sequences increase the endonuclease activity of the Cas9-RNP. (A) Scatterplots of the relative binding affinity (x-axis) *versus* the relative endonuclease activity (y-axis) for all sequences, and **(B-G)** for each library separately. In general, mismatches decrease the affinity and the endonuclease activity toward the lower left. However, mismatches in the PAM-proximal region can often increase the endonuclease activity (y-axis > 0) while decreasing the affinity (particularly apparent in **B-E**). Error bars represent the standard deviation based on two *Spec-seq* and *SEAM-seq* repeats. **(H,I)** Plot of the effect of single mismatches on binding affinity and endonuclease activity. The penalty of single mismatch on affinity is relatively low (< 20-fold), even in the PAM-proximal region. However,

(I) all single mismatches in the PAM-proximal region (1-11) increased endonuclease activity whereas all mismatches in PAM-distal region (12-20) reduced the endonuclease activity.

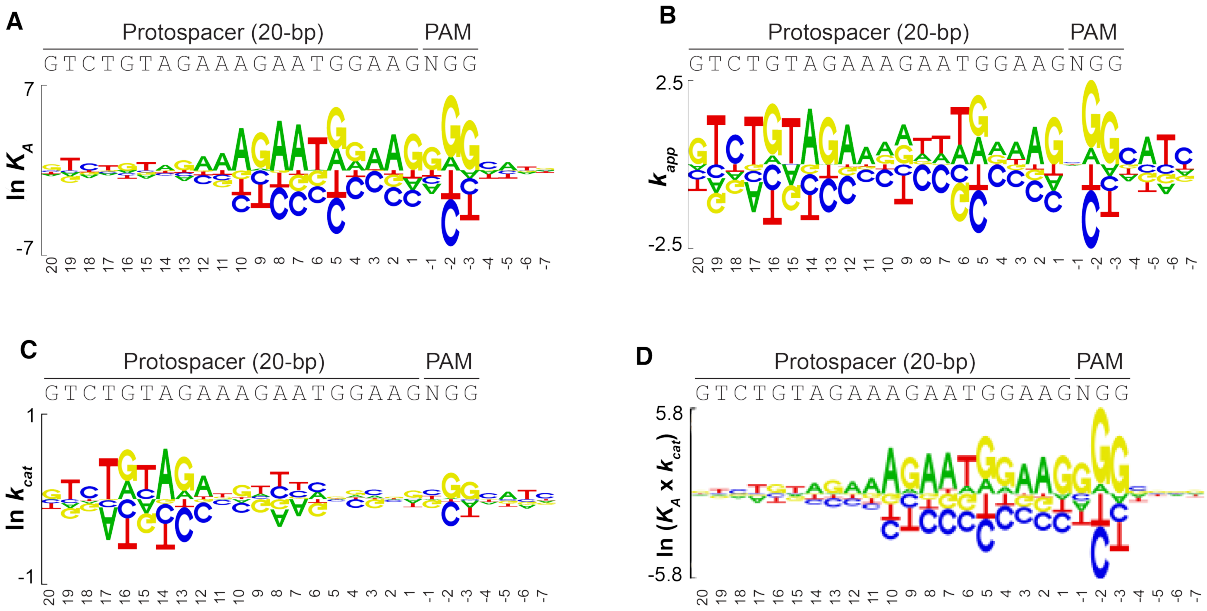


Figure 5. Accurate modeling both binding affinity (K_A) and apparent endonuclease activity (k_{app}). Sequence logos showing how the different bases contribute to the binding affinity **(A)** and apparent endonuclease activity **(B)** using regression models that incorporate the contribution of mononucleotide and dinucleotide features, and corrected for non-specific binding. **(C)** Sequence logo depicting the sequence dependence of endonuclease activity corrected for occupancy activity (k_{cat}) derived using both *SEAM-seq* and *Spec-seq* data. **(D)** The enzymatic efficiency model of Cas9, which is the product of K_A and k_{cat}.

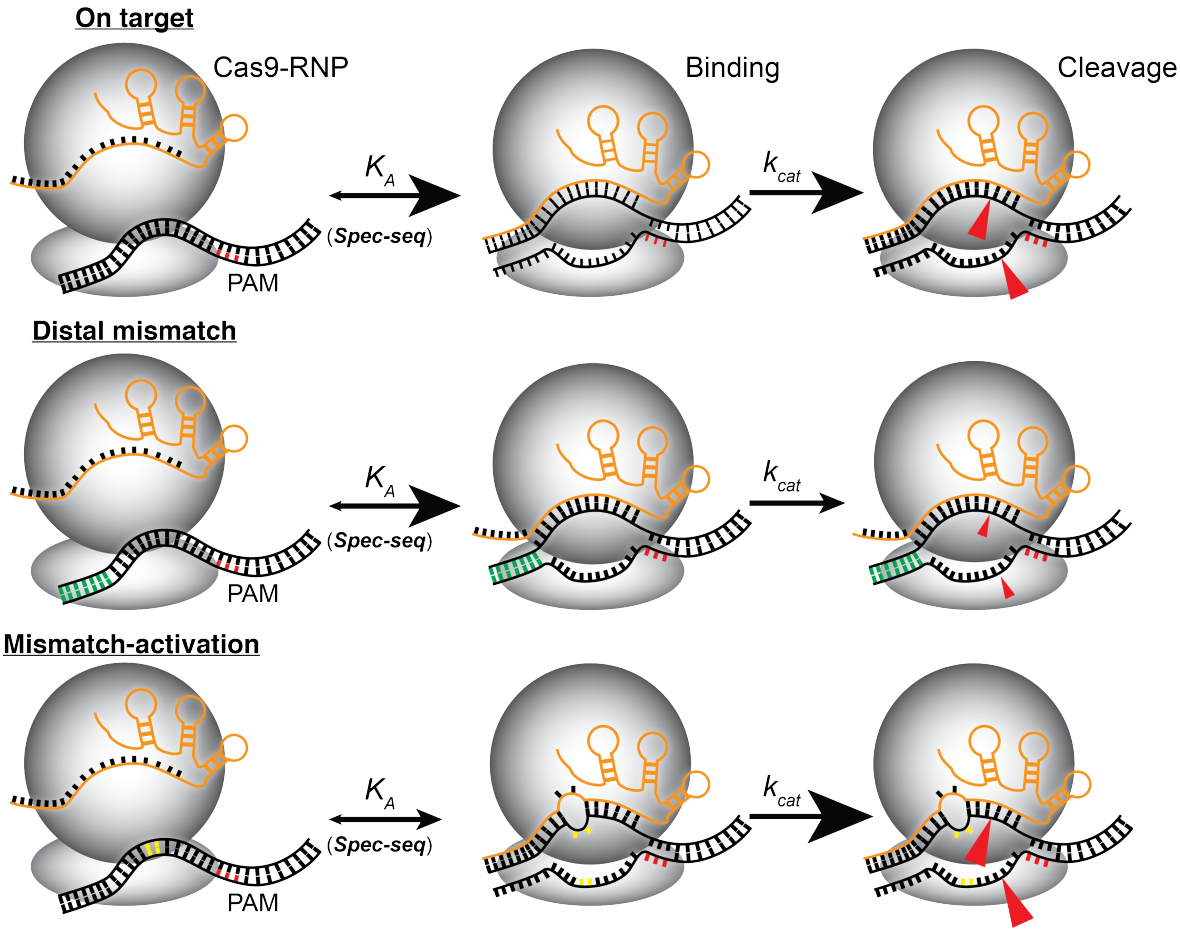


Figure 6. Distinct impact of mismatches on DNA binding and cleavage activity of the Cas9-RNP. Models summarizing the effect of sequence on both affinity (K_A) and turnover (k_{cat}) of the enzyme. The perfect match (top) has both high affinity and activity. Mismatches in the PAM-distal region (green) do not affect affinity, but severely impair k_{cat}. Single and double mismatches within the PAM-proximal region (yellow) can decrease affinity, but substantially increase turnover (k_{cat}).